Selection Models

1 What is the Selection Problem?

There is some confusion as to what the cause of selection issues actually is. So we should begin by outlining the problem up front. I'll provide two examples that come from Sartori (2003, 114) and Achen (1986, 73-76).

1.1 Effect of Education on Women's Wages

Say we want to estimate the effect of education on women's wages. The OLS regression for this would be

$$y_i = x_i\beta + \epsilon_i \tag{1}$$

where y_i is the woman's wage and x_i is her education. The basic selection problem arises in that the sample consists only of women who choose to work and these women may differ in important *unmeasured* ways from women who do not work. For example, women who are smarter may be more likely to enter the labor market. The 'selection equation' for entering the labor market might be:

$$U_i = w_i\gamma + u_i \tag{2}$$

where U_i represents the utility to woman i of entering the labor market and w_i is a vector of factors known to influence a woman's decision to work such as her education level. u_i is assumed to be jointly normally distributed with ϵ_i and contains any unmeasured characteristics in the selection equation. We don't actually observe U_i. All we observe is a dichotomous variable Z_i with a value of 1 if the woman enters the labor force ($U_i > 0$) and 0 otherwise.

So, where does the selection problem actually come from? Well, there are two selection effects.

1. Women with higher levels of education will be more likely to enter the labor force and so we will have a sample of educated women. As Sartori (2003, 114) notes, this non-random aspect of the sample is what is commonly *misunderstood* to be the problem of 'selection bias'. But this on its own does not bias the estimation of the outcome equation in (1).

2. The second selection effect, which is the most important, *is that some uneducated women will go to work.* This is because these women decide that work is worthwhile because they have a high value on some unmeasured variable which is captured in u_i in (3). In other words, these women get into our sample not because they have high education (they have low values of $w_i\gamma$), but because they have large error terms. In contrast, those women who get into our sample because they have high education (large values of $w_i\gamma$) will have a more normal range of errors. The problem is that whether or not education (or independent variables of interest in the outcome equation) is correlated with the unmeasured intelligence (our unmeasured variable) in the overall population, these two variables will be correlated in the selected sample. If intelligence does lead to higher wages, then we will underestimate the effect of education on wages because in the selected sample women with little education are unusually smart.

1.2 Effect of GRE Scores on Grades in Graduate School

Suppose that an admissions committee want to know how GRE scores affect the likelihood of success in graduate school. The problem is that information about success in graduate school (grades) is only available for those students who were admitted. The admissions committee wish to forecast outcomes in the whole pool of applicants but are forced to rely solely on experience with a non-random subset of them. Let's assume that we have the following model. The selection equation for getting admitted might be

$$\text{Admission Rating} = \gamma_0 + \gamma_1 \text{GRE} + u_i \tag{3}$$

$$\text{Admission} = \begin{cases} 1 & \text{if Admission Rating} \geq 0 \\ 0 & \text{if Admission Rating} < 0 \end{cases}$$

where ADMISSION RATING is the latent variable measuring the underlying propensity to be admitted, GRE represents a student's GRE score, and Admission is a dichtomous variable indicating whether the student was admitted or not.

The outcome equation is

$$\text{Success} = \begin{cases} \beta_0 + \beta_1 \text{GRE} + \epsilon_i & \text{if Admission=1} \\ \text{Unobserved} & \text{if Admission} = 0 \end{cases}$$

Admitted graduate students are not representative of applicants generally as the admission equation makes clear. There are many college graduates with low grades who attempt to enroll in graduate school; only a few succeed. These exceptions usually owe their success to favorable (unmeasured) personal characteristics other than grades. While many of these personal characteristics will have no affect on their success in graduate school, it seems reasonable to think that some of them probably will. As a result, there will be some students with low grades who make into graduate school because they have large error terms (they have strong personal characteristics). As a result, this low-grade subset of students will perform above the level of other applicants with the same college grades and so they are no longer representative.

Now suppose the admissions committee examine graduate student grades to compare the performance of those who entered with low GREs to those who entered with high GREs. The group of students who were admitted because they had strong GREs will be representative of the group of applicants with strong GREs. However, the subset of admitted students with low GREs will not be representative of the group of applicants with low GREs - they will perform better in graduate school (because of their large disturbance terms due to personal characteristics) than applicants with low GREs that were not admitted. Ultimately, it may appear that students with high GREs do not outperform students with low GREs in graduate school. The admissions committee might be tempted to conclude that GREs do not predict success. However, intuition makes it clear that this result does not extend to the applicant pool where students with low GREs would, in general, perform quite poorly had they been admitted. In effect, if a random sample of applicants were admitted to graduate school, GREs would be a good predictor of their success.

1.3 The Problem - Omitted Variable Bias

Note that a selection problem does not exist in two types of situations (Achen 1986, 78-79). First, it might be the case that the unmeasured factors influencing the selection equation are uncorrelated with the unmeasured factors influencing the outcome equation. In other words, we would have to assume in our second example that the unmeasured personal characteristics that get a low-GRE student into graduate school are uncorrelated with the unmeasured factors that influence success in graduate school. In effect, we would have to assume that these unmeasured personal characteristics do not influence graduate school success. Is this assumption reasonable? In my opinion, it does not seem particularly reasonable in the GRE example. Second, there is no selection problem if *every* variable influencing selection is controlled in the outcome equation. The problem is that most selection processes are complex and the complete list of variables influencing selection is often not measured, cannot be measured, or unknown. If these two types of situation do not occur, then we have a selection bias problem.

As should be obvious from the two examples given above, the selection bias problem arises because the error term in the outcome equation is correlated with the error term in the selection equation. This means that the error term in the outcome equation will not have mean zero and will be correlated with the explanatory variables. This, in turn, leads to inconsistent estimates. To see this more clearly, recall that Heckman (1979) shows that selection bias is equivalent to an omitted variable bias. For example, he shows in terms of our first example of women's wages above that:

$$E[y_i|U_i > 0] = x_i\beta + \theta \left[\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}\right] \tag{4}$$

where $U_i > 0$ means that the observation was selected into the sample. In other words, this is the expected value of a woman's wage given that the woman is actually working. Note that if we use OLS on the outcome equation in (1), we would be omitting $\theta\left[\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}\right]$. Remember that $\left[\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}\right]$ is the inverse Mill's ratio from last week.

2 An Aside: Incidental Truncation in a Bivariate Distribution

The selection bias problem described above arises due to an incidental truncation of the sample (Greene 2003, 780). A brief description of incidental truncation will make the Heckman model much easier to understand. Suppose that y and z have a bivariate distribution with correlation ρ. We are interested in the distribution of y given that z exceeds a particular value or truncation point τ. Greene (2003, 781) provides the moments for an incidentally-truncated bivariate normal distribution

$$E[y|z > \tau] = \mu_y + \rho\sigma_y\lambda(\alpha_z) \tag{5}$$

$$\text{Var}[y|z > \tau] = \sigma_y^2[1 - \rho^2\delta(\alpha_z)] \tag{6}$$

where $\alpha_z = \frac{\tau - \mu_y}{\sigma_z}$, $\phi(\alpha)$ is the standard normal density,

$$\lambda(\alpha_z) = \frac{\phi(\alpha_z)}{1 - \Phi(\alpha_z)} \tag{7}$$

$$\delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z] \tag{8}$$

$\lambda(\alpha_z)$ is the inverse Mills ratio (IMR) for z.[1] Note that the moments of the incidentally truncated bivariate normal distribution are identical to those of the truncated normal distribution (Greene 2003, 759) with the exception of the ρ terms that we now have. As a result, these equations should look familiar from last week.

[1] As before, these equations all assume that truncation is from below i.e. $z > \tau$. If the truncation is from above ($z < \tau$), then we just make $\lambda(\alpha_z) = \frac{-\phi(\alpha_z)}{\Phi(\alpha_z)}$.

3 Heckman Model

3.1 The Basic Setup

The Heckman model essentially just applies the moments of the incidentally truncated bivariate normal distribution to a data generating process similar to that outlined in the first section of the notes. Let's start with a basic selection equation

$$z_i^* = w_i\gamma + u_i \tag{9}$$

$$z_i = \begin{cases} 1 & \text{if } z_i^* > 0 \\ 0 & \text{if } z_i^* \leq 0 \end{cases}$$

and a basic outcome equation

$$y_i = \begin{cases} x_i\beta + \epsilon_i & \text{if } z_i^* > 0 \\ - & \text{if } z_i^* \leq 0 \end{cases}$$

As noted above, problems arise when estimating β if u_i and ϵ_i are correlated. Note that it should now be obvious that the tobit model is the special case where $y_i = z_i$. Typically, we also make the following assumption about the distribution of, and relationship between, the error terms in the selection and outcome equation:

$$\begin{aligned} u_i &\sim N(0,1) \\ \epsilon_i &\sim N(0,\sigma^2) \\ \text{corr}(u_i,\epsilon_i) &= \rho \end{aligned} \tag{10}$$

Put differently, we typically assume a bivariate normal distribution with zero means and correlation ρ. There is no generally accepted name for this model. Amemiya (1985, 384) calls it a 'type 2 Tobit model'. Others call it a 'generalized tobit model' or simply the 'sample selection model'. We'll refer to it as the Heckman (1979) model.

3.2 Conditional Means in the Heckman Model

Now that we have our basic setup, all we have to do is insert these into the relevant equations for the moments of the incidentally truncated bivariate normal distribution given earlier. Thus,

$$\begin{aligned} E[y_i|y_i \text{ is observed}] &= E[y_i|z_i^* > 0] \\ &= E[x_i\beta + \epsilon_i|w_i\gamma + u_i > 0] \\ &= x_i\beta + E[\epsilon_i|w_i\gamma + u_i > 0] \\ &= x_i\beta + E[\epsilon_i|u_i > -w_i\gamma] \end{aligned} \tag{11}$$

If the errors ϵ_i and u_i are independent, then the last term simplifies to E[ϵ_i]=0 and OLS regression of y_i on x_i will give consistent estimates of β. However, any correlation between the two errors means that the truncated mean is no longer $x_i\beta$ and we need to take account of selection. Thus, we need to obtain $E[\epsilon_i|u_i > -w_i\gamma]$ when ϵ_i and u_i are correlated. As Greene (2003, 782) notes,

$$E[\epsilon_i|u_i > -w_i\gamma] = \rho\sigma_\epsilon\lambda_i(\alpha_u) \tag{12}$$

where $\alpha_u = \frac{-w_i\gamma}{\sigma_u}$, $\lambda(\alpha_u) = \frac{\phi(\frac{-w_i\gamma}{\sigma_u})}{1-\Phi(\frac{-w_i\gamma}{\sigma_u})} = \frac{\phi(\frac{w_i\gamma}{\sigma_u})}{\Phi(\frac{w_i\gamma}{\sigma_u})}$. Thus, the conditional mean in the Heckman model (bivariate selection model) is:[2]

$$\begin{aligned} E[y_i|y_i \text{ is observed}] &= E[y_i|z_i^* > 0] \\ &= E[x_i\beta + \epsilon_i|w_i\gamma + u_i > 0] \\ &= x_i\beta + E[\epsilon_i|w_i\gamma + u_i > 0] \\ &= x_i\beta + E[\epsilon_i|u_i > -w_i\gamma] \\ &= x_i\beta + \rho\sigma_\epsilon \left[\frac{\phi(\frac{w_i\gamma}{\sigma_u})}{\Phi(\frac{w_i\gamma}{\sigma_u})}\right] \\ &= x_i\beta + \rho\sigma_\epsilon\lambda_i(\alpha_u) \\ &= x_i\beta + \beta_\lambda\lambda_i(\alpha_u) \end{aligned} \tag{15}$$

Note that this is where our earlier equation (4) came from. Thus, we now have

$$\begin{aligned} y_i|z_i^* > 0 &= E[y_i|z_i^* > 0] + \nu_i \\ &= x_i\beta + \beta_\lambda\lambda_i(\alpha_u) + \nu_i \end{aligned} \tag{16}$$

Again, this clearly illustrates that OLS on just the outcome equation would lead to biased and inconsistent estimates because $\beta_\lambda(\alpha_u)$ is omitted. Note that the variance equation in (6) implies that even if $\lambda_i(\alpha_u)$ were included in the model, OLS would be inefficient since ν_i is heteroskedastic.

3.3 Marginal Effects

So, what are the marginal effects in the Heckman model? The marginal effect of the independent variables on y_i in the observed sample consists of two components.[3] First, there is the direct effect of the independent variable on the mean of y_i which is captured by β. Second, there is an indirect effect if the independent variable also appears in the selection equation. This is because a change in some x not only changes the mean of y_i, but also the probability that an observation is actually in the sample i.e. it will affect y_i through $\lambda(\alpha_u)$. Thus, the marginal effect of x on y_i in the observed sample is

$$\frac{\partial E[y_i|z_i^* > 0]}{\partial x_{ik}} = \beta_k - \gamma_k \left(\frac{\rho\sigma_\epsilon}{\sigma_u}\right)\delta_i(\alpha_u) \tag{17}$$

[2] It is interesting to note the connection with the tobit model. In the Heckman model we have

$$E[y_i|z_i^* > 0] = x_i\beta + \rho\sigma_\epsilon\lambda_i(\alpha_u) \tag{13}$$

In the tobit model from last week, we have

$$E[y|y > 0] = x_i\beta + \sigma\lambda(\alpha) \tag{14}$$

As you can see, the main difference is that there is a separate latent variable doing the censoring in the Heckman model that is different to the variable determining the outcome equation. This difference also requires that we take account of the correlation between the disturbances in the selection and outcome equations i.e. the ρ.

[3] Note that this is typically the quantity of interest. We might occasionally be interested in the marginal effect of variables for the whole sample - those observed in the sample and those not observed in the sample. However, you would need to use a different equation for this.

where $\delta_i(\alpha_u) = [\lambda_i(\alpha_u)]^2 - \alpha_u \lambda_i(\alpha_u)$.[4]

The equation for the marginal effect in (19) is important to note. The main point is that if $\rho \neq 0$ and the independent variable appears in the selection and outcome equation, then β_k does NOT indicate the marginal effect of x_k on y_i. As Greene (2003, 783) notes, it is quite possible for the magnitude, sign, and statistical significance of the marginal effect to all be different from the estimate of β_k. This point is often ignored. Thus, it is not sufficient to simply estimate the model and look at t-statistics to know if an independent variable (that appears in the selection and outcome equation) has an effect on y_i.

3.4 Estimation

There are two-ways of estimating the Heckman model.

1. **Heckman's Two-Step Procedure**

   In this model, all we need to do is assume that u_i and ϵ_i are independent of the explanatory variables with mean zero and that $u_i \sim N(0, 1)$ (Wooldridge 2002, 562). The two-step procedure is the most common method for estimating the Heckman model and is as follows:

   (a) Estimate the probit equation (selection equation) by MLE to obtain estimates of γ. For each observation in the selected sample, compute $\hat{\lambda}_i = \frac{\phi(w_i\hat{\gamma})}{\Phi(w_i\hat{\gamma})}$ (the inverse Mills ratio) and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i - w_i\hat{\gamma})$.[5]

   (b) Estimate β and $\beta_\lambda = \rho\sigma_\epsilon$ by OLS of y on x and $\hat{\lambda}$.

   The estimators from this two-step procedure are consistent and asymptotically normal. This procedure is often called a 'Heckit model'.

[4] Note that σ_u is assumed to be 1 if we use Probit to estimate the selection equation. Thus, the equations that we have been looking at can be simplified somewhat (see Greene (2003, 784). For example, the conditional mean would be

$$\begin{aligned} E[y_i|y_i \text{ is observed}] &= E[y_i|z_i^* > 0] \\ &= E[x_i\beta + \epsilon_i|w_i\gamma + u_i > 0] \\ &= x_i\beta + E[\epsilon_i|w_i\gamma + u_i > 0] \\ &= x_i\beta + E[\epsilon_i|u_i > -w_i\gamma] \\ &= x_i\beta + \rho\sigma_\epsilon\left[\frac{\phi(\frac{w_i\gamma}{\sigma_u})}{\Phi(\frac{w_i\gamma}{\sigma_u})}\right] \\ &= x_i\beta + \rho\sigma_\epsilon\left[\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}\right] \\ &= x_i\beta + \beta_\lambda\lambda_i(\alpha_u) \qquad (18) \end{aligned}$$

where we now have $\alpha_u = -w_i\gamma$ and $\lambda(\alpha_u) = \frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}$. The marginal effect would just be

$$\frac{\partial E[y_i|z_i^* > 0]}{\partial x_{ik}} = \beta_k - \gamma_k(\rho\sigma_\epsilon)\delta_i(\alpha_u) \qquad (19)$$

[5] The $\hat{\delta}_i$ bit is useful for obtaining correct standard errors in the second stage. Recall that it was part of the variance for an incidentally truncated bivariate normal distribution.

2. **MLE Version**

   The Heckman model can also be estimated by MLE. However, to do so requires making a stronger assumption than those required for the two-step procedure. For MLE, we need to assume that u_i and ϵ_i are distributed bivariate normal with mean zero, that $u_i \sim N(0, 1)$, that $\epsilon_i \sim N(0, \sigma^2)$, and that corr$(u_i, \epsilon_i) = \rho$. Thus, the MLE estimation is not as general as the two-step procedure. As Wooldridge (2002, 566) notes, another drawback is that this procedure is less robust than the two-step procedure and is sometimes difficult to get it to converge. However, MLE estimation will be more efficient if u_i and ϵ_i really are jointly normally distributed. To see the loglikelihood for the MLE version, see the appendix in chapter 5 of Berinsky (2004) or the STATA manual.

3.4.1 Standard Errors

Note that the standard errors in the outcome equation will need to be corrected. There are two reasons for this. The first is that we have heteroskedasticity if $\beta_\lambda \neq 0$ (i.e. if we have selection bias). We could easily correct for this by using robust standard errors. However, there is a second reason why our errors need to be corrected. $\hat{\gamma}$ is an estimator of γ. This means that $\hat{\lambda}$ is also just an estimator of λ. We need to take account of the fact that our inverse Mills ratio is estimated with uncertainty. This second reason means that we cannot just simply use robust standard errors - we need to do something more complicated. To see a discussion of how the standard errors are corrected, see Heckman (1979), Wooldridge (2002, 564) and Greene (2003, 785). STATA will automatically correct the standard errors for you.

3.4.2 Identification

Note that, technically, the Heckman model is identified when the same independent variables in the selection equation appear in the outcome equation. However, identification only occurs on the basis of distributional assumptions about the residuals alone and not due to variation in the explanatory variables. Identification is essentially possible due to the non-linearity in the selection equation (the non-linearity is introduced into the outcome equation through the inverse Mill's ratio). However, this is not the only problem that arises if you have all of the variables from the selection equation in the outcome equation. You will probably get imprecise estimates in the outcome equation. In fact, if there is little variation in $w_i\gamma$ in the selection equation, then $\hat{\lambda}$ can be approximated quite well by a linear function of w. If this is the case, you will have difficulty getting precise estimates in the outcome equation because you will have high multicollinearity and large standard errors. This is the case even if you don't include all of the variables from the selection equation in the outcome equation. Basically, the point is that if the selection equation is not very good at determining selection, then you are likely to get imprecise estimates in the outcome equation.

Because of issues with identification, we nearly always want at least one independent variable that appears in the selection equation but does not appear in the outcome equation i.e. we need a variable that affects selection, but not the outcome (Sartori 2003, 112). As you might expect, it may be difficult to think of a variable that affects selection but not the outcome. But this is what you need to do! Many people will either just drop a variable from the outcome equation or add a variable to the selection equation – however, this is often theoretically unmotivated and means that the outcome or selection equation is incorrectly specified. One solution is that presented by Sartori (2003).

3.4.3 Selection Bias

The coefficient on the inverse Mill's ratio will indicate if there is selection bias. If the coefficient is statistically significant, then we know that there was selection bias. However, it is somewhat hard to say much if the coefficient is not significant. We can say that there is no selection bias as we have formulated the selection equation. However, this is assuming that we have the selection equation correct.

4 Selection Bias in Binary Choice Models

The Heckman model is useful for handling linear regression models when there is a selection mechanism at work i.e. when the outcome equation involves a continuous dependent variable. However, we are often interested in cases where the outcome equation involves a dichotomous dependent variable. In effect, we would have a probit selection equation and a probit outcome equation.[6] Neal Beck calls this a 'double probit model' but it appears that few others do. As we'll see, most people call it a 'bivariate probit model with selection' and this corresponds to the 'heckprob' command in STATA. Before we get to the bivariate probit model with selection, let's start by looking at the straightforward bivariate probit model.

4.1 Bivariate Probit

In the bivariate probit model, we have two separate probit models with correlated disturbances in the same spirit as the seemingly unrelated regression models (SUR).[7] In effect, we have two binary dependent variables, y_j, $j = 1, 2$. These often represent two interrelated decisions by some actor. For example, Staton (2006) uses a bivariate probit model because he is interested in modeling a court's decision (i) to invalidate policies that are challenged and (ii) to issue a press release about its decision. You could also think of a decision to vote for a House and Senate candidate on the same ballot or a state's decision to adopt two different, but related, policy initiatives. The basic idea is that the two decisions are interrelated. Thus, we might have the following model

$$\begin{aligned} y_1^* &= x_1\beta_1 + \epsilon_1 \\ y_2^* &= x_2\beta_2 + \epsilon_2 \end{aligned}$$

where y_j^* are unobservable and related to the binary dependent variables y_j by the following rule

$$y_j = \begin{cases} 1 & \text{if } y_j^* > 0 \\ 0 & \text{if } y_j^* \leq 0 \end{cases}$$

for j=1, 2. If the errors between the two probit models are independent of one another i.e. Cov$[\epsilon_1, \epsilon_2] = 0$, then we can just estimate the two probit models separately.

But, what happens if we have:

$$\begin{aligned} \epsilon_{1i} &= \eta_i + u_{1i} \\ \epsilon_{2i} &= \eta_i + u_{2i} \end{aligned}$$

In other words, the errors in each model consist of a part (u_i) that is unique to that model and a second part (η_i) that is common to both - the error terms are correlated. We might assume that all three types of errors

[6] You could, of course, use logit, but it seems that virtually everyone uses probit.

[7] Much of this section comes from notes by Zorn.

are normally distributed. If this is true, then the ϵ_is will also be normal but they will also be dependent. This means that each ϵ_i now depends, in part, on the value of η_i, and this in turn means that ϵ_{1i} and ϵ_{2i} will be related to one another.

We should care about this because it makes a difference.[8] We're interested in

$$\begin{aligned} Pr(y_{1i} = 1) &= Pr(\epsilon_{1i} > -x_1\beta_1) \\ &= Pr(u_{1i} + \eta_i > -x_1\beta_1) \end{aligned} \tag{20}$$

and

$$\begin{aligned} Pr(y_{2i} = 1) &= Pr(\epsilon_{2i} > -x_1\beta_2) \\ &= Pr(u_{2i} + \eta_i > -x_2\beta_2) \end{aligned} \tag{21}$$

In other words, we're interested in the joint probability of y_1 and y_2.

We know that if two random variables are independent, then their joint probability is just the product of their marginal probabilities. So, if y_1 and y_2 are independent, then we would have:

$$\begin{aligned} Pr(y_1 = 1, y_2 = 1) &= F(y_1) \times F(y_2) \\ Pr(y_1 = 1, y_2 = 0) &= F(y_1) \times [1 - F(y_2)] \\ Pr(y_1 = 0, y_2 = 1) &= [1 - F(y_1)] \times F(y_2) \\ Pr(y_1 = 0, y_2 = 0) &= [1 - F(y_1)] \times [1 - F(y_2)] \end{aligned} \tag{22}$$

The problem is that in our situation, the two probabilities are not independent since they both depend on the common value of η_i. So, now we need to calculate joint probabilities for non-independent events. Recall from your probability classes that

$$\begin{aligned} Pr(A\text{and}B) &= Pr(A|B) \times Pr(B) \\ &= Pr(A) \times Pr(B|A) \end{aligned} \tag{23}$$

It follows from this that

$$\begin{aligned} Pr(y_1 = 1, y_2 = 1) &= Pr(y_1 = 1|y_2 = 1) \times Pr(y_2 = 1) \\ &= Pr(y_1 = 1) \times Pr(y_2 = 1|y_1 = 1) \end{aligned} \tag{24}$$

To get at this, we need to assume some *joint distribution* for the y_is. When the y_is are dependent, we pick some *bivariate joint distribution*. As you'll probably have realized, we typically use a bivariate normal distribution. For two standard-normally distributed ϵs, their joint pdf will be

$$\phi_2 = \phi(\epsilon_1, \epsilon_2) = \frac{1}{2\pi\sigma_{\epsilon_1}\sigma_{\epsilon_2}\sqrt{1-\rho^2}}\text{exp}\left[-\frac{1}{2}\left(\frac{\epsilon_1^2 + \epsilon_2^2 - 2\rho\epsilon_1\epsilon_2}{1-\rho^2}\right)\right] \tag{25}$$

where ρ is a correlation parameter denoting the extent to which the two ϵs covary. Their joint cdf will be

$$\Phi_2 = \Phi(\epsilon_1, \epsilon_2) = \int_{\epsilon_1}\int_{\epsilon_2} \phi_2(\epsilon_1, \epsilon_2, \rho)d\epsilon_1 d\epsilon_2 \tag{26}$$

[8] It is possible to estimate two probit equations separately and obtain consistent results. However, when $\rho \neq 0$, it is more efficient to estimate the equations jointly.

If $\rho = 0$, then the two variables (or errors) are independent and the Φ_2 reduces to two separate standard normal distributions. If $\rho \neq 0$, then the two variables (errors) are correlated and the probability of one variable will be dependent on the value/probability of the other. If $\rho = 1$, then the two variables are essentially (exactly) the same. If $\rho = -1$, then the two are exactly negatively related i.e. their scales are reversed.

We use the Φ_2 distribution to estimate bivariate probit models. In other words, we typically assume that the errors are independent and identically distributed as a standard bivariate normal with correlation ρ i.e.

$$\begin{aligned} E[\epsilon_1|x_1,x_2] &= E[\epsilon_2|x_1,x_2] = 0 \\ \text{Var}[\epsilon_1|x_1,x_2] &= \text{Var}[\epsilon_2|x_1,x_2] = 0 \\ \text{Cov}[\epsilon_1,\epsilon_2|x_1,x_2] &= \rho \end{aligned}$$

This is often written as $\epsilon_{1i}, \epsilon_{2i} \sim \phi_2(0,0,1,1,\rho)$. Given this, we can now make probability statements about y_i. For example,

$$\begin{aligned} Pr(y_{1i}=1, y_{2i}=1) &= \int_{-\infty}^{\epsilon_{1i}} \int_{-\infty}^{\epsilon_{2i}} \phi_2(x_1\beta_1, x_2\beta_2;\rho) d\epsilon_{1i} d\epsilon_{2i} \\ &= \Phi_2(x_1\beta_1, x_2\beta_2;\rho) \end{aligned} \tag{27}$$

4.2 Estimation

As in the standard probit model, each observation contributes some combination of $Pr(y_k = 1)$ for $k \in 1,2$ depending on their specific values on those variables. In other words, the log-likelihood is just a sum across the four possible transition probabilities (i.e. the four possible combinations of y_1 and y_2) multiplied by their associated probabilities.

The log-likelihood for the bivariate probit model is

$$\begin{aligned} \ln L = \sum_{i=1}^{N} \{ & y_{i1}y_{i2}\ln\Phi_2(x_1\beta_1, x_2\beta_2;\rho) \\ & + y_{i1}(1-y_{i2})\ln[\Phi(x_1\beta_1) - \Phi_2(x_1\beta_1, x_2\beta_2;\rho)] \\ & + (1-y_{i1})y_{i2}\ln[\Phi(x_2\beta_2) - \Phi_2(x_1\beta_1, x_2\beta_2;\rho)] \\ & + (1-y_{i1})(1-y_{i2})\ln[1 - \Phi(x_1\beta_1) - \Phi(x_2\beta_2) - \Phi_2(x_1\beta_1, x_2\beta_2;\rho)]\} \end{aligned} \tag{28}$$

where $\Phi_2(\cdot,\cdot;\rho)$ denotes the bivariate standard normal cdf with correlation coefficient ρ and $\Phi(\cdot)$ is the univariate standard normal cdf we have seen before. Just to clarify, $\Phi(x_1\beta_1) - \Phi_2(x_1\beta_1, x_2\beta_2;\rho)$ is just the probability that $y_1 = 1$ minus the probability that $y_1 = y_2 = 1$. In other words, it captures Pr($y_1 = 1$, $y_2 = 0$).

To estimate this model in STATA, you have two options. You can type:

- biprobit y_1 y_2 x_1 x_2 x_3 $\ldots$

if you have the same independent variables in each probit model. Or, you can estimated a seemingly unrelated version of the bivariate probit model by typing

- biprobit (y_1 = x_{11} x_{12} x_{13} $\ldots$) (y_2 = x_{21} x_{22} x_{23} $\ldots$)

This allows you to have different independent variables in each of the probit models. The models are the same.

4.3 Interpretation

4.3.1 Model Fit

Two separate probits are ‘nested’ in the bivariate probit model since they occur if $\rho = 0$. Thus, we can test the hypothesis that the bivariate probit model fits the data better than the separate probits by conducting a likelihood ratio test. For the separate probits, the joint likelihood is just the product of the two separate (marginal) likelihoods. This means that the joint log-likelihood is just the sum of the two log-likelihoods. We can compare this joint log-likelihood of the separate models to that for the bivariate probit model using a standard LR test. You can also calculate predicted probabilities and compare them to actual outcomes as we did in previous weeks when we looked at probit etc..

4.3.2 Marginal Effects

We can look at the coefficients and their standard errors to determine the direction and statistical significance of the individual variables. There are several different expected values that you might be interested in. You may be interested in the unconditional mean function:

$$\begin{aligned} E[y_1] &= \Phi(x_1\beta_1) \\ E[y_2] &= \Phi(x_2\beta_2) \end{aligned} \tag{29}$$

You could then look at marginal effects by taking the derivative of these with respect to some independent variable of interest. This is exactly the same situation as when we discussed the univariate probit models earlier in the semester.

However, you may also be interested in the conditional mean function:

$$\begin{aligned} E[y_1|y_2 = 1] = Pr(y_1 = 1|y_2 = 1) &= \frac{Pr(y_1 = 1, y_2 = 1)}{Pr(y_2 = 1)} \\ &= \frac{\Phi_2(x_1\beta_1, x_2\beta_2; \rho)}{\Phi(x_2\beta_2)} \end{aligned} \tag{30}$$

You could take the derivative of this with respect to some independent variable to get the marginal effects (see Greene (2003, 713)).

4.3.3 Predicted Probabilities

Probably better is to calculate predicted probabilities or changes in predicted probabilities;

$$\begin{aligned} Pr(y_1 = 1, y_2 = 1) &= \Phi_2(x_1\hat{\beta}_1, x_2\hat{\beta}_2; \hat{\rho}) \\ Pr(y_1 = 1, y_2 = 0) &= \Phi(x_1\hat{\beta}_1) - \Phi_2(x_1\hat{\beta}_1, x_2\hat{\beta}_2; \hat{\rho}) \\ Pr(y_1 = 0, y_2 = 1) &= \Phi(x_2\hat{\beta}_2) - \Phi_2(x_1\hat{\beta}_1, x_2\hat{\beta}_2; \hat{\rho}) \\ Pr(y_1 = 0, y_2 = 0) &= 1 - \Phi(x_1\hat{\beta}_1) - \Phi(x_2\hat{\beta}_2) - \Phi_2(x_1\hat{\beta}_1, x_2\hat{\beta}_2; \hat{\rho})\} \end{aligned} \tag{31}$$

You can get STATA to produce predicted probabilities by typing

- predict pr11, p11

4.4 Some Conclusions

Note that whenever $\hat{\rho} \neq 0$, variables in one of the probit models have an indirect effect on the other probit model. It can sometimes be hard to get the model to converge. Model specification is very important. As Zorn notes, if the true model for $Pr(y_1 = 1, y_2 = 1)$ depends on x_1, x_2, and x_3, and you omit x_3, then your error terms are now

$$\begin{aligned}\epsilon_{1i} &= u_{1i} + \beta_3 x_3 \\ \epsilon_{2i} &= u_{2i} + \beta_3 x_3\end{aligned}$$

which are correlated by construction. This means that an estimate of $\hat{\rho} \neq 0$ may be due to actual correlation between the two processes or due to simple specification error.

5 Bivariate Probit with Partial Observability

In the bivariate probit model we have just looked at, we observe both y_1 and y_2.[9] However, Poirier (1980) developed what he called a 'bivariate probit model with partial observability'. In this model, we do not observe y_1 and y_2 in all circumstances. Consider a scenario where two people on a committee vote anonymously on a motion under a unanimity rule. If we observe the motion pass, then we know that $y_1 = 1$ and $y_2 = 1$. However, if the motion does not pass, we do not know whether this is because (i) $y_1 = 1$ and $y_2 = 0$, (ii) $y_1 = 0$ and $y_2 = 1$, or (iii) $y_1 = 0$ and $y_2 = 0$. In effect, instead of observing x_i, y_{1i}, and y_{2i}, we observe only x_i and z_i, where $z_i = y_{i1} \times y_{i2}$. Thus, we have

$$z_i = \begin{cases} 1 & \text{if } y_{i1} = 1 \text{ and } y_{i2} = 1 \\ 0 & \text{otherwise} \end{cases}$$

In terms of the four possible voting combinations – 'yes, yes', 'no, yes', 'yes, no', and 'no, no' – the last three are indistinguishable because all we can observe for these outcomes is that the motion does not pass.

Przeworski and Vreeland (2002) employ a variant of the bivariate probit model with partial observability to study bilateral cooperation between national governments and the International Monetary Fund. IMF agreements are only observed when both the national government and the IMF agree to them. Not observing an IMF agreement may be because either one side or the other does not want to participate, or because neither side wants to participate.

5.1 Estimation

It turns out that there are several variants of the bivariate probit model with partial observability. We will look at two.

1. **Poirier's (1980) Model**

   In the Poirier model, y_1 and y_2 are simultaneously determined and their two errors (ϵ_1 and ϵ_2) are correlated. Thus,

$$\begin{aligned}Pr(z = 1) &= \Phi_2(x_1\beta_1, x_2\beta_2; \rho) \\ Pr(z = 0) &= 1 - \Phi_2(x_1\beta_1, x_2\beta_2; \rho)\end{aligned}$$

[9] For a good discussion of observability in bivariate probit models, see Meng and Schmidt (1985).

It follows directly form this that the log-likelihood function for this version of the bivariate probit model with partial observability is:

$$\ln L = \sum_{i=1}^{N} \{z_i \ln\Phi_2(x_1\beta_1, x_2\beta_2; \rho) + (1 - z_i)\ln[1 - \Phi_2(x_1\beta_1, x_2\beta_2; \rho)]\} \tag{32}$$

where $\Phi_2(x_1\beta_1, x_2\beta_2; \rho) = Pr(y_{1i} = 1, y_{2i} = 1) = Pr(z_i = 1)$.

This is the bivariate probit model with partial observability that you can estimate in STATA. To see how, consider the data set in Table 1. With partial observability, we know only that 14 outcomes are positive for both y_1 and y_2. Thus, you'll have a variable z that has 14 observations coded as 1 and 60 observations coded as 0. So, how do you estimate this model? You need to use STATA's 'biprobit' command. However, to use the biprobit command, you need two dependent variables - this is a problem since you only have one i.e. $z = y_1 \times y_2$. All you need to do is to create another variable that is identical to z, say z_2 and add this into the model as well. Thus, you would type:

- generate z2=z
- biprobit, z z2 x1 x2 x3 . . . , partial

You can use the same commands as before to put in different variables for each of the individual actors that determine z.

Table 1: Possible Data for Bivariate Probit Model with Partial Observability

		y_2		
		0	1	Total
y_1	0	26	26	52
	1	8	14	22
	Total	34	40	74

2. **Abowd and Farber's (1982) Model**

In this model, y_1 and y_2 are determined sequentially and the errors (ϵ_1 and ϵ_2) are assumed uncorrelated. Abowd and Farber's example is of an individual who decides to enter a queue and then subsequently decides whether or not to accept an offer upon reaching his turn in the queue. This is the version of the bivariate probit model with partial observability used by Przeworski and Vreeland (2002). From this setup we have:

$$\begin{aligned} Pr(z = 1) &= \Phi(x_1\beta_1)\Phi(x_2\beta_2) \\ Pr(z = 0) &= 1 - [\Phi(x_1\beta_1)\Phi(x_2\beta_2)] \end{aligned}$$

It follows from this that the log-likelihood function is just

$$\ln L = \sum_{i=1}^{N} \{z_i \ln\Phi(x_1\beta_1)\Phi(x_2\beta_2) + (1 - z_i)\ln[1 - \Phi(x_1\beta_1)\Phi(x_2\beta_2)]\} \tag{33}$$

This version of the bivariate probit model with partial observability is not canned in STATA but can be estimated in LIMDEP.

6 Bivariate Probit with Sample Selection

This is the equivalent of Heckman's selection model except now we have a probit model in the selection equation and a probit model in the outcome equation. This model has various names - 'double probit', 'bivariate probit with sample selection', 'censored probit', 'bivariate probit with partial partial observability'. This model is somewhere between a bivariate probit model and a bivariate probit model with partial observability. For example, we observe more than in the partial observability model but less than in the full bivariate probit model. In the terminology that we have been using, we observe x_i, y_{i1} and $z_i = y_{i1} \times y_{i2}$. Essentially, we observe y_2 if and only if $y_1 = 1$ (this is so because if $y_{i1} = 1$, then $y_{i2} = z_i$ and z_i is observed. However, if $y_{i1} = 0$, then we have no information about y_{i2}). Thus, the first probit equation is completely observed, but we have only a selected (censored) sample for the second. Note that in terms of the four possible outcomes, two ($y_{i1} = 0$, $y_{i2} = 1$ and $y_{i1} = 0, y_{i2} = 0$) are indistinguishable. This is an improvement in observability relative to the bivariate probit model with partial observability, which had three indistinguishable outcomes.

Berinsky (2004) provides an example of this model when he examines attitudes towards race issues. He argues that he has a selected sample since he only observes race attitudes for those people willing to answer the question in the first place. Thus, the first probit model is whether the respondent answers the question and the second probit model is their dichotomous attitude towards race issues.

6.1 Estimation

There are three types of observations in a sample with the following probabilities.

$$\begin{array}{ll}
y_1 = 0 & Pr(y_1 = 0) = \Phi(-x_1\beta_1) \\
y_1 = 1, y_2 = 0 & Pr(y_1 = 1, y_2 = 0) = \Phi(x_1\beta_1) - \Phi_2(x_1\beta_1, x_2\beta_2, \rho) \\
y_1 = 1, y_2 = 1 & Pr(y_1 = 1, y_2 = 1) = \Phi_2(x_1\beta_1, x_2\beta_2, \rho)
\end{array}$$

From this, it is easy to generate the log-likelihood function

$$\begin{aligned}
\text{ln}L = \sum_{i=1}^{N} \{ & y_{i1}y_{i2}\text{ln}\Phi_2(x_1\beta_1, x_2\beta_2; \rho) \\
& + y_{i1}(1 - y_{i2})\text{ln}[\Phi(x_1\beta_1) - \Phi_2(x_1\beta_1, x_2\beta_2; \rho)] \\
& + (1 - y_{i1})\text{ln}\Phi(-x_1\beta_1)\}
\end{aligned} \tag{34}$$

As with the Heckman model, you'd want to include at least one variable in the selection equation that does not appear in the outcome equation. To estimate this model in STATA, you type:

- heckprob y_2 x_{21} x_{22} x_{23} $\ldots$, sel(y_1=x_{11} x_{12} x_{13})

References

Abowd, John M. & Henry S. Farber. 1982. "Job Queues and the Union Status of Workers." *Industrial and Labor Relations Review* 35:354–367.

Achen, Christopher. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge: Harvard University Press.

Berinsky, Adam. 2004. *Silent Voices: Opinion Polls and Political Representation in America*. Princeton: Princeton University Press.

Greene, William. 2003. *Econometric Analysis*. New Jersey: Prentice Hall.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–161.

Meng, Chun-Lo & Peter Schmidt. 1985. "On the Cost of Partial Observability in the Bivariate Probit Model." *International Economic Review* 26:71–85.

Poirier, D. J. 1980. "Partial Observability in Bivariate Probit Models." *Journal of Econometrics* 12:210–217.

Przeworski, Adam & James Raymond Vreeland. 2002. "A Statistical Model of Bilateral Cooperation." *Political Analysis* 10:101–112.

Sartori, Anna. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11:111–138.

Staton, Jeffrey K. 2006. "Constitutional Review and the Selective Promotion of Case Results." *American Journal of Political Science* 50:98–112.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.